Zone Analysis in Biology Articles as a Basis for Information Extraction

Yoko MIZUTA [a] Anna KORHONEN [a] Tony MULLEN [a]
Nigel COLLIER [a,*]

[a] *National Institute of Informatics*
2-1-2 Hitotsubashi Chiyoda-ku
101-8430 Tokyo, Japan

Abstract

In the field of biomedicine, an overwhelming amount of experimental data has become available as a result of the high throughput of research in this domain. The amount of results reported has now grown beyond the limits of what can be managed by manual means. This makes it increasingly difficult for the researchers in this area to keep up with the latest developments. Information extraction (IE) in the biological domain aims to provide an effective automatic means to dynamically manage the information contained in archived journal articles and abstract collections and thus help researchers in their work. However, while considerable advances have been made in certain areas of IE, pinpointing and organizing factual information (such as experimental results) remains a challenge. In this paper we propose tackling this task by incorporating into IE information about rhetorical zones, i.e. classification of spans of text in terms of argumentation and intellectual attribution. As the first step towards this goal, we introduce a scheme for annotating biological texts for rhetorical zones and provide a qualitative and quantitative analysis of the data annotated according to this scheme. We also discuss our preliminary research on automatic zone analysis, and its incorporation into our IE framework.

Key words: Bioinformatics, Journal article, Classification, Multiple classification analysis

* Corresponding author.
Email addresses: ymizuta@nii.ac.jp (Yoko MIZUTA), korhonen@nii.ac.jp (Anna KORHONEN), mullen@nii.ac.jp (Tony MULLEN), collier@nii.ac.jp (Nigel COLLIER).

1 Introduction

Information extraction (IE) in the biological domain is now regarded as an essential technique for utilizing information contained in archived journal articles and abstract collections such as MEDLINE. Major domain databases offer large-scale archives of semi-structured results, but they tend to be incomplete or not up-to-date. For the latest results, as well as for confirmation of results reported in the database and for supplementary information, access to the latest literature is necessary. Thus, the significance of a more sensible management of facts, specifically an integration and update of experimental results, is self-evident. Given the limitations of manual work for such purposes in terms of both efficiency and accuracy, IE's focus on factual information is of a critical importance.

Recent intensive research in natural language processing in the biological domain (bioNLP) has made major progress in the extraction of bio-named entities and biological interactions (e.g. [1], [2], [3]), but further advancement aimed at pinpointing and organizing factual information remains a challenge. In particular, the important task of identifying new experimental results is complicated by the large number of statements made in each article that pertain to results in general, including references to previous work as well as technical details and conjectures. The same information (about a molecular event, for example) may be provided as a new result, as a previously known result, as a conjecture, etc., that is, in different rhetorical contexts. Because current IE relies on surface lexical and syntactic patterns, it is not sensitive to the rhetorical status of information. As a consequence, existing techniques tend to extract old results mixed with new ones, leaving the novel contribution unclear. We expect that a rhetorical analysis of biological texts and its incorporation into IE will provide a means to address this problem.

As the first step towards this goal, we proposed in [4] annotating biology texts in terms of rhetorical zones with a shallow nesting using a scheme modified from [5]. In [6], we gave a qualitative analysis of our sample hand-annotated data and described how such zones can be identified by humans. In this paper, we provide a comprehensive qualitative and quantitative analysis of the process and the results of zone analysis (ZA) in the hand-annotated data, describe some preliminary work on automatic ZA and discuss its future incorporation into our framework for IE.

The organization of the paper is as follows. We first discuss in more detail the motivations for ZA in biology, from an IE perspective and in the view of previous work (Section 2). We then introduce our framework. We describe general characteristics of biological articles and introduce our annotation scheme and annotated data (Section 3). We then shed light on the decision process

in which zones can be identified by a human annotator and describe, mostly qualitatively, the main features required for identification of each zone class (Section 4). To summarize the results of the annotation work, we illustrate the distribution of zones in the annotated articles both visually and quantitatively (Section 5). Finally, we discuss ongoing and future work on automatic ZA and its integration to IE (Section 6).

2 The need for ZA in biomedicine

2.1 Critical issues in bioNLP

We discuss below the critical issues in bioNLP involved in pin-pointing and organizing factual information and describe how ZA can be applied to help this process.

First, we argue that the current IE techniques do not effectively allow us to distinguish between different kinds of factual information, although input data contains material sufficient for this purpose. In particular, biological articles include useful information about various rhetorical statuses (classification of text in terms of argumentation and intellectual attribution), such as new vs. old results; the author's own work vs. somebody else's work, as illustrated below (dots indicate deleted content words; boldface is for emphasis):

(1) **Recent data suggest that** ..... is involved in DPC removal in mammalian cells (ref.). **The data presented here suggest that** at least some subset of DPCs can be removed by NER mechanism .... (PNAS3, p.1908) [1]

The two sentences above are both interpreted as making a biological statement. However, there is a crucial difference between the two statements. The first statement is based on previous data provided in the literature, whereas the second is based on the author's own data provided in the present paper. Automatically preprocessing the text in terms of the rhetorical status of information prior to the application of IE should help to extract information about new results in articles (as exemplified in the second statement above).

Second, so far the scope of IE in biology has largely focused on abstracts. Arguably, however, the goal should be full texts, given their much richer sources of information and the increasing ease of access (e.g. on-line journal portals

[1] Henceforth, the source of an example is shown using the artcie ID (in this case 'PNAS3'). Details of source journal articles are provided in the 'Source online journal articles' section toward the end of this paper.

and collections such as PUBMED-central). Focusing on full articles, however, requires exploring new IE techniques because full articles are linguistically more challenging than abstracts. For example, full texts present much more complexity in the sentence structure and vocabulary (e.g. inserted phrases, embedded sentences, nominalization of verbs, anaphoric expressions). Their analysis requires a much more complex set of patterns and algorithms than those available in existing IE systems. One solution to this problem is to identify the parts of the article relevant to the information we are interested in before attempting a more complex analysis. This is where rhetorical analysis can help. For example, if we want to extract information about certain kinds of biological interactions found by the author, we could focus on the statements describing the author's own experimental results and findings and ignore all other statements in the text (e.g. an overview of previous work or technical details of the experiments).

Third, because current IE techniques basically rely on rules based on surface syntactic and morphological forms they do not provide the means to extract information about experimental results in an organized manner. However, automatic means for organizing such data would be important because experimental results only make full sense in their relation to other information such as the experimental goal and procedure. ZA could help to address this problem, as well as to help correctly identify demonstrative and anaphoric references in the texts (e.g. *these results*; *it*).

From these points of view, we expect that ZA can play an essential role in bioNLP aiming to improve IE.

2.2 Approaches to rhetorical analysis

Previous work on rhetorical analysis of scientific articles can be classified roughly into the following categories;

Genre analysis The analysis of the text structure and linguistic properties found in each genre: e.g. [7], [8], [9].

Discourse analysis The analysis of discourse relations in a hierarchical structure. e.g. Rhetorical Structure Theory (RST) by [10] for an analysis in terms of conceptually-defined relations (e.g. CAUSE, SOLUTION and INTERPRETATION) and Penn Discourse Tree Bank (PDTB) by [11] for an analysis in terms of discourse connectives (e.g. *because* and *therefore*).

Argumentative zoning The analysis of the global rhetorical status of each sentence or other constituent in the text, dividing the text into zones in a flat structure. e.g. Work by [5] in terms of zones such as OWN for the author's own work and OTHER for somebody else's work.

While we share (and are aware of) the insights of both genre analysis and discourse analysis, our work fits best into the last category, since we focus on the type of information which is global to the article. For example, in our ZA, reference to previous work as 'background information' remains as such whether it is supported or refuted by the author later in the article. This difference, irrelevant to our work, is fundamental to the other two lines of approach (e.g. SUPPORT vs. ANTITHESIS relations in RST).

In the Appendix, a sample illustration of RST analysis is provided in comparison with the zone analysis in our framework. The illustration shows the type of information important to our task and explains why no detailed hierarchical structure is needed in our work. In our work, part of the information about the discourse, especially, how the argumentation develops, is exhibited in the pattern (in terms of order and combination) in which zones appear in texts.

2.3 Zoning by Teufel et al.

Our approach is partly based on the work by [5]. Their work focuses on automatic text summarization of computer science articles. They propose analyzing the text into rhetorical zones in a flat structure in terms of argumentation and intellectual attribution. They offer a total of seven zone classes including AIM (the aim of the present work), BACKGROUND (general background), OTHER (other people's work), and OWN (the author's own work described in the present paper). [5] report that their zones can be to a large extent automatically detected using ML, and demonstrate the role of zoning in improving the quality of text summarization.

Our current focus is on the biological domain, while we eventually aim at a generalization to a larger scientific domain to the extent that this is possible. We propose our scheme from both domain-specific and general perspectives. For one thing, the biological domain deserves particular attention in its own right, given the pressing need for bioNLP. For another thing, the biological domain demonstrates homogeneity in comparison with e.g. the domain of computer science, as indicated by the articles we investigated. Thus we consider that it makes sense to take advantage of such domain-specific characteristics in designing our scheme. In what follows, we aim to distinguish between domain-specific adjustments and general improvements, applicable also to other domains including computer science.

3 Framework

For the reasons mentioned above, we have made some major modifications to the original scheme of [5]. These are for conceptual clarification and for a closer look at the author's own work focusing on the experimental results. In what follows, we first describe the general characteristics of biological articles and then introduce our annotation scheme in terms of zone classes and the principles of annotation.

3.1 General characteristics of biological articles

The biological journals we investigated (see Section 3.5 for details) have a rigid section format, consisting of an "Abstract" and the "Introduction", "Materials and Methods", "Results", and "Discussions" sections (henceforth, the I-, M-, R-, and D-sections, respectively). In what follows, we focus on the four main sections, excluding the "Abstract" [2].

The articles investigated fit into an experimental framework and each section exemplifies a typical pattern of argumentation. The whole article is committed to some main problem-solving task such as 'the identification of the effect of Mnt deletion in governing key Myc functions' or 'the identification of the receptor for MAG that elicit morphological changes in neurons'. The main task is divided into smaller problem-solving tasks.

We can generalize as follows [3]:

- The I-section introduces the main problem to be solved and outlines the content of an article.
- The M-section states methodological details.
- The R-section states smaller problem-solving units in terms of experimental procedure and results (and their interpretation).
- The D-section synthesizes the results and findings and provides summarizing/prospective remarks.

Importantly, factual information is provided across sections; from a broader perspective in the I- and the D-sections, and from a more specific perspective in the M- and the R-sections [4]. Thus, we cannot simply choose a specific section from which to extract the information we need.

[2] Some articles, usually shorter ones, have the R- and the D-sections in the combined format. Here we focus on the standard 4-section format.
[3] For details of text structure, see for example [9] and [12].
[4] See Section 5 for the distribution of information in terms of zones.

Our zone classes are designed to enable our system to extract the type of information to our concern. The major originalities in our scheme are the following.

(1) Zone classes are divided into three groups on the semantic basis [5]. This is for conceptual clarity in the annotation task, and for future representational purposes when we use appropriate software.

We consider that the proposed grouping applies across different scientific domains.

(2) A fine-grained OWN class is proposed to distinguish between different aspects of the author's own work.

We consider that a fine-grained OWN class is a general necessity, assuming that most space in any article is devoted to describing the author's own work [6]. The specific set of subclasses, however, may vary depending on the domain. The subclasses we propose in our scheme are designed for the biological articles investigated, describing work conducted in an experimental framework. In those articles we found various descriptions pertaining to the author's work such as methodological details, concrete experimental results, a smaller number of statements about the interpretations of the results, etc. From the IE perspective, it is important to distinguish between such different kinds of information.

(3) CNN (Connection) and DFF (Difference) classes are defined to capture statements about the relations between data and/or findings provided by the author or by somebody else. These are extensions of the BASIS and CONTRAST classes in [5]; BASIS is for the statements about other work as a basis for the present work and CONTRAST is for the statements contrasting the author's work with other work.

This is a domain-specific adjustment. In the domain of computer science, which [13] investigated, the focus of comparison is on showing the stance and significance of the author's work relative to work by others. In contrast, in the domain of biology, where the methodology is rather established, we observed that the focus is on a more neutral comparison between the author's data/findings and those by others.

Our complete set of zone groups and classes is shown in Table 1.

[5] This grouping is an advancement from the version proposed in [6] .

[6] [5] report that 67% of the whole sentences in their corpus are annotated as OWN.

Group 1 This group concerns major elements with respect to the problem-solving process, intellectual attribution, and scientific argumentation. The description of the zone classes belonging to this group follows.

BKG (Background) Given information (reference to previous work; general assumptions).

PBM (Problem-setting) A problem or an open issue which the author identifies or introduces, and which motivates the author's work presented in the paper. This is typically the goal of the present research/paper or the goal of a specific experiment performed.

OWN The author's own work in the following aspects:

- MTH (Method): Statements about the experimental procedure and the materials used.
- RSL (Result): Statements about data in observed experimental results.
- INS (Insight): The author's interpretation of the data as it pertains to the experimental goal. It concerns a biological process or the role of a biological entity behind the observed results, and other findings of a biological significance.
- IMP (Implication): Various kinds of implications of the author's work described in the present paper, which would motivate a new experimental paradigm. Typical cases are; assessment, applications, limitations of the present work, and future work. This class also covers the author's conjectures and hypotheses.
- ELS (Else): Any other kind of information within OWN (e.g. naming statements).

Group 2 This group deals with comparative or contrasting relations between items which fit into Group 1 classes. Specifically, similarities or differences are described between results, insights, etc. presented in the work at hand and between items pertaining to the work at hand and those pertaining to previous work.

CNN (Connection) Correlation, consistency.

DFF (Difference) Contrast, inconsistency.

Group 3 This group concerns statements about the paper/work at hand. It consists of one zone class only.

OTL (Outline) A characterization or summary of the paper; excerpts from the paper; statements about the section organization.

Table 1
Zone groups and classes employed in our scheme

3.3 Unit of annotation

The granularity of annotation is inherently a controversial issue. On the one hand, we need to cover the semantically-motivated zone classes of interest as proposed above. On the other hand, we need to obtain a high level of (inter- and intra-)annotator agreement and to make the task simpler in terms of the

number of steps required. For the former purpose (i.e. a better coverage), a finer level of annotation would be appropriate, whereas for the latter purpose (i.e. simplicity and annotator agreement), a coarser annotation would be better [7]. As a trade-off, we decided to take the following approach [8].

In general, the annotator proceeds sentence by sentence. If a sentence fits semantically into a single zone class, it is annotated as a zone of this class. If adjacent sentences fit into the same class, they may be annotated either together or separately. In both cases, each component sentence comes with the context of the same zone class.

In some cases, a single sentence consists of multiple constituents fitting into different zones. Certain types of constituents qualify for an independent zone. They are variants of clauses, and *to*-phrases. This is because a rhetorical status apparently corresponds to a proposition, the closest syntactic counterpart of which is a clause [9]. We also find this practical for the following reasons. Our earlier proposal was to use a sentence as a unit of annotation. However, we observed many cases where a single sentence consists of constituents (e.g. clauses) pertaining to different rhetorical statuses. In such a case, giving a single annotation to the whole sentence would be a problem, because it is not trivial at all for the annotator to identify the 'most important/relevant' zone class for the sentence, as observed during our annotation task. Thus, a sentence-level annotation would result in an inter-annotator disagreement (i.e. disagreement between multiple people engaged in annotation). Therefore we decided on the clause-level annotation.

Propositions, however, have surface syntactic variations, as illustrated in the following sentence: 'Reducing the amount of X resulted in Y' (X and Y indicate unspecified content words). Here the subject 'Reducing the amount of X' expresses an event which the author was engaged in, whereas Y expresses the resultant state or event. Thus both of these semantically correspond to a proposition, although they are noun phrases instead of sentences. Therefore we consider that it is necessary to specify in syntactic terms the constituent types which qualify for an independent zone in our scheme. Otherwise we may introduce too many levels of annotation, leading to too complex and detailed results (e.g. at the noun phrase level). The final list of such constituent types is provided in Table 2.

[7] For relevant discussions on the unit of annotation made in different frameworks, see for example [14] in RST and [11] in PDTB.

[8] For the annotation guidelines written from a more practical perspective, see [15].

[9] We share this view with [14] and [11].

(1) A sequence of sentences

(2) A sentence

(3) Coordinate clauses
e.g. A but B. ⇒ [A] [but B.]
Note: Other coordinate conjuncts include *and*, *whereas*, and *while* (with the contrast meaning).

(4) Subordinate clauses
e.g. A when B. ⇒ [A] [when B.]; When B, A ⇒ [When B,] [A.]
Note: Other subordinate conjuncts include *because*, *since*, *although* and *when*.

(5) Nonrestrictive relative clauses
e.g. [. . . . . , [which . . . ,] . . . . .]
Note: The inserted clause constitutes an embedded zone.

(6) Present or past participle version of nonrestrictive relative clauses
e.g. [. . . . . ,] [indicating that . . . .] cf. [. . . ,] [which indicates that . . .]

(7) *To*-infinitives expressing the goal/ purpose or the result of what's stated in the remainder of the sentence
e.g. [To test a subset of these candidates, . . .] [we performed . . . .]

Table 2
Constituent types qualifying for an independent zone in our scheme (square brackets in the examples indicate zone boundaries.)

3.4 Nested annotation

We observed that the sentence structure in full texts is rather complicated, and that constituents are often nested. Typical cases within a sentence are relative clauses and noun modification. In other cases multiple sentences as a whole fit into a single zone class yet some parts of it fit into different zone classes. Therefore, we consider that nested annotation is necessary, even though it complicates annotation. We allow for a 2-level nesting (not more) within each zone group.

We also find the need for multiple annotation of a single unit. Empirical analysis indicates that even though zone classes are conceptually non-overlapping, an annotation unit may fit into multiple classes. That is, a surface linguistic unit such as a sentence and a clause may well represent complex concepts. This is because there is a discrepancy between a syntactic/morphological and a semantic unit. The following example illustrates complex zones motivating multiple annotation (dots indicate deleted content words).

(2) Similar DNA kinks were also observed in the complexes with . . . (ref.), which show structural similarities with Dna domain IV . . . (NAR6, p.2083)

Sentence (2) both provides a result and compares it with other results. Thus, the sentence fits into RSL (result) and CNN (connection) zones simultaneously. It is a case of combined zones, which is conceptually distinct from indeterminacy between two zones. (In the case of indeterminacy, only one answer is expected: The question is what is *the* answer.)

Combined zones are precisely speaking different from nested zones. The former concern the combination of zones at the same level, whereas the latter concern embedding of zones. However, currently we annotate combined zones as nested zones sharing an identical scope. Combined zones interpreted in this sense are not sensitive to the ordering of component zones (in this case CNN and RSL). Thus the zone annotation of (2) may be either of the following:

(3) a. [[Similar DNA kinks were also observed ..... $]_{CNN}$ $]_{RSL}$
b. [[Similar DNA kinks were also observed ..... $]_{RSL}$ $]_{CNN}$

3.5 Annotation data

We created our hand-annotation data from a total of 20 articles randomly selected from those published online between 2002 and 2003 in four major online journals (EMBO: European Molecular Biology Organization, NAR: Nucleic Acid Research, PNAS: Proceeding of National Academy of Science, and JCB: Journal of Cell Biology).

A single annotator (i.e. the first author of the present paper) was engaged in all the hand-annotation. Annotation task and the design/revision of the annotation scheme gave feedback to each other. That is, the first article was annotated using the original scheme proposed by [5], and some modifications on the scheme were made based on the requirements from our IE perspective. Then another article was annotated using the revised scheme and further elaborations were made, and so forth.

This iterative approach works out because an annotation scheme should be designed both on a theoretical and an empirical basis. For example, there were several potential ways of defining the fine-grained OWN classes or the unit of annotation (e.g. sentence, clause). An ideal solution can only be identified after actually looking at the articles in the domain under study and doing the annotation task. After having worked on several articles, however, we came up with a rather clear idea about our annotation scheme.

At the moment we haven't used a dedicated tool for the annotation task. We downloaded the articles as Word documents and annotated them by manually inserting start/end tags in XML-format. The annotated files were then converted into text format. After a clean-up process (e.g. correction of mistyped

labels) the data was processed further for statistical analysis to identify, for example, the distribution of zones and useful features related to the identification of zones. We call the resulting dataset the ZAISA-1 dataset.

4 Main features of each zone

This section describes the feature types and the decision process used by the human annotator and the main characteristics of each zone based on our annotation task and data.

4.1 Decision tree for a human annotator

Figure 1 provides the decision tree used for annotation by a human. For the correct annotation of any constituent (as illustrated in Table 2), it is important to consider a larger context in which the constituent appears. For example, the description of an existing problem can be annotated as PBM (Problem-setting) only when the problem is relevant as the motivation for the author's present work. Otherwise it should be annotated as BKG instead.

4.2 Feature types

The list below shows the main feature types which we found helpful in our annotation task. The contribution of each feature type varies from one zone class to another. The features are illustrated further in the subsequent sections.

- Lexical: words and phrases with a specific meaning.
  e.g.1. We *report* the results of our experiments ...
  e.g.2. ... *remains unclear*.
- Main verb: the verb playing the central role in a sentence/constituent. In the case of an embedded sentence, this is the verb in the main clause.
- Tense (past, present, and present perfect)
- Modal auxiliaries (e.g. *would*, *might*, *can*)
- Section (I, M, R, D) and the location within the section or a paragraph (e.g. beginning, middle, end)
- Sentence-final citation (having the whole sentence as its scope), including self-citation of previous or ongoing work by the author
- Reference to Figure or Table representing the author's results
- Underlying subject of the sentence
  e.g. *We* performed an experiment using ....; *The amount of X* increased.

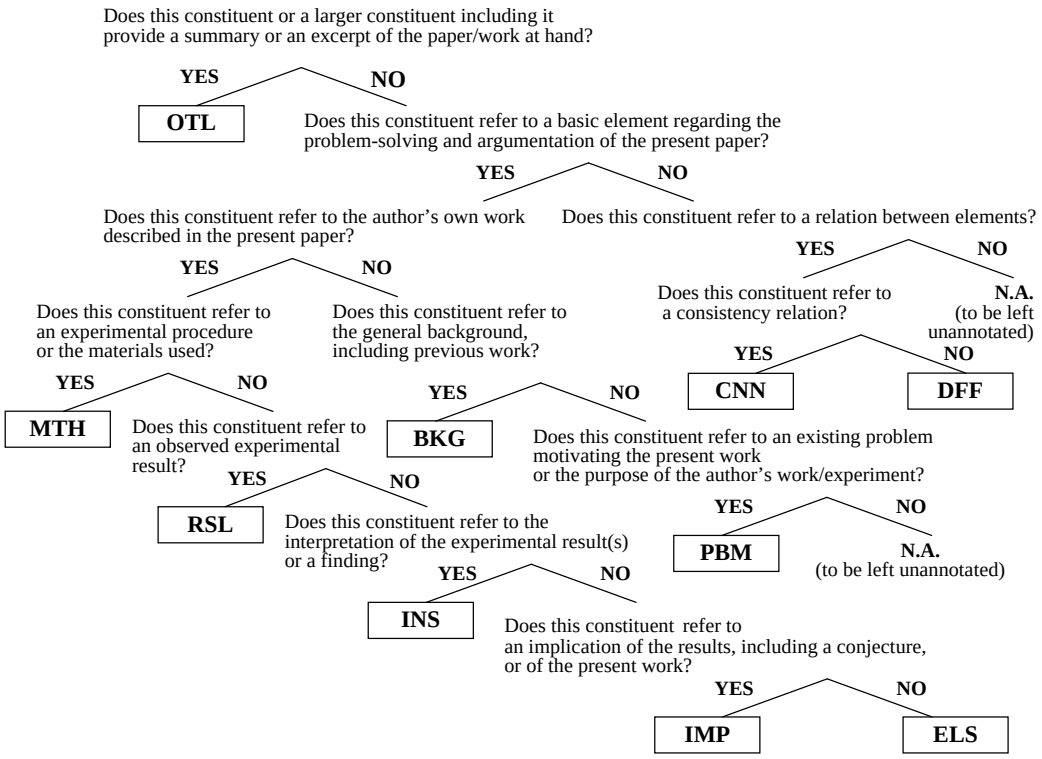

Fig. 1. Decision tree for zone annotation

4.3 Qualitative analysis of each zone class

Below we discuss and analyze each zone class in our scheme with respect to the features used for identifying the class, starting from the predominating feature where possible.

4.3.1 OUTLINE (OTL)

OTL provides a concise characterization or a summary of the present work. The zone typically appears either toward the end of the I-section or at the beginning or the ending of the D-section. In other words, it appears before or after the main body of the paper.

(4) We report here the results of experiments ..... In brief, we have asked, ... and ..... To address the first question, we utilized ..... We found ..... Together, these results not only confirm that ... but also that ... (End of the I-section) (NAR1, pp.1830-1831)

The above example starts with the description of the main focus of the paper (the first sentence) followed by 'excerpts' from the paper (the rest). All the elements are embedded in a reporting context and provide abstract-like information. We consider that the whole sequence of sentences deserve an independent class from both theoretical and practical perspectives. Thus, we propose introducing the OTL class.

The beginning of an OTL zone often would be an AIM zone employed in the scheme of [5], and includes certain kinds of linguistic cues:

(5) a. Indexicals. e.g. in *this paper*; in *the present study*; *here*
    b. Verbs used for the author's presentational/reporting purposes
    e.g. (we) *show*; *demonstrate*; *present*; *report*

As for excerpts from the paper, each element also fits into other zones (in Group 1 or 2) in its own right. In the above example, the second sentences through the last fit respectively into PBM, MTH, RSL, and INS in Group 1.

The tenses used in OTL are typically simple present or future (in the I-section), and present perfect or simple past (in the D-section).

4.3.2 BACKGROUND (BKG)

Below are some examples of elements fitting into a BKG zone.

(6) a. In cells, DNA is tightly associated with a variety of proteins that serve both to maintain the structural organization of the genetic material and to coordinate cellular processes including replication, repair, recombination, and transcription. (PNAS3, p.1905)
    b. Increasing evidence (Ref.) suggests that there is a link between anabolic energy metabolism and appetite control ..... It is well documented (Ref.) that fatty acid synthesis in lipogenic tissues, ..., occurs only during energy surplus ..... (PNAS4, p.1921)
    c. Analyses of data generated in cell culture revealed the existence of ...and suggested that more than one repair pathway can be involved in the repair of these lesions (Ref.). (PNAS3, p.1905)
    d. A wide variety of restriction-modification (R-M) systems have been discovered and characterised. (NAR3, p.1888)

In the I-section, the majority of BKG elements appear in long unbroken sequences (see Tables 5 and 6 in Section 5 for illustration of statistical distributions). This is intuitive, given the introductory nature of this section.

Taken in isolation, there are no strong features for BKG zones, except for a sentence-final citation. As shown in the above examples, BKG presents three

tense variations:

- Simple present for a generic statement about background information (biological facts and previous work)
- Simple past (to mention previous work)
- Present perfect, to mention previous work relevant to the present situation

A wide range of verbs are used to cover both biological and bibliographical facts. Sentence-final citations having as their scope the whole sentence signal BKG, but inter-sentential citations having a smaller scope (e.g. citations referring to a named entity) do not.

In sections other than the I-section, BKG zones are rather dispersed. They can often be identified by virtue of lacking clear signals for other zones.

4.3.3 PROBLEM SETTING (PBM)

There are two types of PBM zones, as illustrated in the examples below.

(7) a. ...has not been {established/ addressed}.
b. There has been no study on ....
c. Little is currently known about ....
d. There are very limited data concerning ....
e. ...remains unclear.

(8) To test {this hypothesis/ whether ...}, ...
To evaluate ..., ...
To address the question of ..., ...

The first PBM type as in (7) shows up in the I-section. It highlights what's lacking in previous research (e.g. knowledge, study, a research question) which motivates the work described in the present paper. Vocabulary with 'negative polarity', expressing certain kinds of negation or incompleteness, appearing in the I-section is the strongest feature for identification of this zone type. Others include tense variation in either simple present or present perfect, depending on the range of time in focus.

The second type of PBM as in (8) is observed in the R-section and represents the goal of a specific experiment. As illustrated in (8), it typically exemplifies in a *to*-phrase appearing sentence-initially (common) and sentence-finally (occasional).

These two types of PBM describe the goal of the research at different levels: The first type concerns the whole work while the second type its subset (i.e. a specific experiment).

4.3.4 METHOD (MTH)

The following are some examples fitting into a MTH zone (italics are ours).

(9) a. We *performed* ..., using .....
b. We *exploited* the observation that .....(Ref.)
c. Next, we *utilized* sucrose-gradient fractionation .... (NAR1, p.1835)
d. ... *was normalized.*

As illustrated statistically in Table 5 in Section 5, MTH is the predominant zone of the M-section. However, also in the R-section, MTH plays an important role together with RSL. The strongest signal for MTH is the semantic type of the main verb of a sentence. The main verbs expressing experimental procedure are relatively more frequent in this zone than on the average. For example in the sample of our 20 articles the main verbs *perform*, *experiment*, *stain*, *wash* and *measure* have, respectively, 78%, 64%, 70%, 100% and 100% of all their occurrences in the MTH zone. Also, the underlying (semantic) subject type 'we' helps to distinguish MTH from RSL (which are often adjacent to each other) in active sentences. In the past tense, MTH typically takes the form of an event description. It shows up either in a passive or an active form. In both cases, the semantic subject is 'we' (as an actor). This is an important difference from passive sentences in RSL zones such as (11a).

A paragraph in the R-section starts usually with a combination of PBM and MTH, as illustrated below [10]:

(10) [To test this hypothesis,]$_{PBM}$ [we performed ....]$_{MTH}$

We observed that when a sentence takes the above form, PBM tends to occur before MTH. This can be explained in terms of the linguistic phenomenon of 'iconicity', the fact that conceptual and/or the real world ordering of elements is often reflected in linguistic expressions. For example, in (10), the PBM portion (*to*-phrase) is preposed in agreement with the fact that the author had the experimental goal before performing the experiment.

4.3.5 RESULT (RSL)

Below are some examples fitting into a RSL zone (italics are ours).

(11) a. Furthermore, the distribution of that signal *was shifted* significantly away from the membrane pellets, and toward the soluble fraction. (NAR1, p.1835)

[10] We illustrate a zone annotation by a pair of square brackets identifying the scope and the immediately following tag indicating the zone class such as PBM.

b. No significant change *was seen* in ....
c. As illustrated ..., cells devoid of Scp160p (striped bars) *demonstrated* a marked enrichment of both ...and ..... (NAR1, p.1835)

Like MTH, RSL describes events in the past tense. The main verb is typically one of those expressing the following:

(1) Observations (e.g. *observe*, *recognize* and *see*, having 'we' as its underlying subject)
(2) Phenomena (e.g. *represent*, *show* and *demonstrate*, having as its subject the material used)
(3) Biological processes (e.g. *mutate*, *translate*, *express*, having as its subject a biological entity, often in the passive form)

A main verb fitting into one of these semantic classes is the strongest feature for the identification of RSL. As with MTH, many typical verb senses occur statistically more frequently in this zone than in other zones. In the R-section, MTH zone is often followed by a RSL zone (each consisting of a sequence of sentences) without any connectives in between. The boundary of these zones can be identified by virtue of the change in the semantic class of the main verb. Unlike MTH, RSL may also be written in the present tense to create a context in which the author observes and presents the results in real time, referring to Figures and Tables.

In the R-section, RSL zones follow MTH with no discourse connectives. The boundary is identified by virtue of a cause-effect relation between the two. Specifically, the main verbs used in these zones play a critical role; some of them present a rather complementary distribution (see the preceding subsection for comparison). In some cases, constituents which semantically fit into MTH and RSL may be combined in a single sentence by the predicate *resulted in* as follows [11]:

(12) Parallel reverse transcription reactions using total RNA isolated from whole cell soluble lysates of both strains *resulted in* indistinguishable strong smears (data not shown). (NAR1, p.1834)

However, the above usage relating a method and the results is not frequent. *Result in* is more commonly used to relate biological events as in (13) below. Also, the explicit use of the noun *result(s)* is uncommon.

(13) Furthermore, loss of Scp160p *resulted in* a significant change in both the abundance and distribution between soluble and membrane-associated fractions .... (NAR1, p.1830)

[11] These constituents, however, do not qualify for independent zones in our scheme, since they do not meet the requirements given in Table 2.

Therefore, keyword searches using *result(s)* do not work out for the purpose of identifying experimental results. In contrast, RSL zones can be identified by using features such as the main verbs and the location. Thus, annotating RSL zones is critical for identifying the result(s) as well as the reference (i.e. content) of *these results* and the like appearing in the text.

(14) Interestingly/ Surprisingly/ Noticeably/ ...

In a RSL zone, empathetic expressions as above may be used to call the reader's attention, often sentence-initially. The adjective version (e.g. *striking*) is also used.

4.3.6 INSIGHT (INS)

We have identified three major signals for the identification of INS. The examples below illustrate the first one (annotations and italics are ours):

(15) [As can be seen in Figure 2C, Z was not significantly different compared with that in Figure 2A,]$_{RSL}$ [*indicating that* ...had no appreciable effect on ....]$_{INS}$ (NAR3, p.1891)

(16) [Interestingly, centrosomal ZYG-9 was significantly reduced in tbg-1(RNAi) embryos (Fig. 2, B and C), to ...levels (...) In the converse experiment, normal levels of γ-tubulin staining were observed in embryos .....]$_{RSL}$ [*These results suggest that* γ-tublin is required to assemble centrosomes .....]$_{INS}$ (JCB2, p.595)

These are conventionalized forms which the authors use to state their interpretation of the data in the observed results with respect to some biological process in concern. A generalization of the pattern is this:

(17) X indicate that Y. (a variant: X, indicating that Y.)
X: results/experiments/studies, Y: biological statement or model
Verb variations found in our sample:
indicate, *suggest*, *demonstrate*, *represent*, *reveal*.

From our empirical analysis of the data, it seems that these particular forms are used to express a direct, objective interpretation (Y) of the data in experimental results (X) as opposed to the author's more subjective interpretations and conjectures which leave room for further investigations in a new experimental paradigm. The special status of the above-mentioned forms conforms to the structural pattern in which they appear. Characteristically, they are observed in the R-section immediately following a RSL zone (in the case of the variant *X, indicating Y*, the interpretation-statement part *Y* is combined with the result-statement part *X* within the same sentence), whereas the au-

thor's more subjective interpretations usually appear in the D-section. Thus, we are motivated to distinguish the first type of interpretation of the data (fitting into INS) from other kinds of interpretations by the author (fitting into IMP discussed below).

The second signal is a sentence containing the main verb *seem/ appear* or *consider*:

(18) a. ...{seem/ appear} to ...
It {seems/ appears} that ...
b. ...is considered to ...

Although these expressions may cover the author's interpretations of more subjective nature, they seem to indicate a higher-level certainty on the part of the author. Thus we are motivated to distinguish the information provided from conjectures etc.

The third signal is the use of verbs such as *confirm* and *support*:

(19) This was confirmed, as shown in Figure 3.
(Note: *This* refers to the author's hypothesis.)

Although (19) refers to the figure which shows the result, the main focus of the sentence is on the author's interpretation of the result. Upon confirmation, the hypothesis has changed into a finding. Thus the sentence fits into INS rather than into RSL.

A generalization of the pattern is this:

(20) X confirm that Y; Y was confirmed.
X: results/experiments/studies
Y: proposition (hypothesis or prediction).

As we will discuss later, *confirm* also signals CNN, relating two things (X and Y above). Therefore, it triggers multiple annotation for INS and CNN.

4.3.7 IMPLICATION (IMP)

The IMP class is used as a cover category for the author's 'weaker' insights obtained from experimental results and for other kinds of implication of the work (e.g. assessment, applications, future work).

(21) Fusion of the Mod and Res subunits of type III enzymes, ..., *would* probably result in type IIG enzymes .... (NAR3, p.1894)

(22) Therefore, ..., we *speculate* that the inactive or interfering amino acid(s) at position 246, ..., induces disordered topology and interferes with .... (PNAS5, p.1824)

'Weaker' insights (vs. 'stronger' insights fitting into INS) are signaled by modal expressions (e.g. *would*, *could*, *may*, *might*, *be possible*, *one possibility is that*) and verbs related to conjecture (e.g. *speculate*, *hypothesize*), as in the examples above.

Assessment and future work are signaled by weak linguistic clues such as the following, respectively:

(23) a. These data are *significant* because ...
b. This approach has the *potential* to increase ...
c. ...provides structural *insights* into ...

(24) a. Potential targets also remain to be studied.
b. We do not yet know ...
c. Further experiments will focus on ....
d. a future study/work/challenge

Taken out of context, IMP mentioning future work looks very similar to (the first type of) PBM mentioning the problems in previous work, unless it contains key words such as *future* and *further*. The critical feature which helps distinguish between IMP and PBM is the section in which they appear: The I-section signals PBM, whereas the D-section signals IMP.

4.3.8 ELSE (ELS)

We found only six instances of ELS in our data. Interestingly, all these instances turned out to be naming statements such as the following:

(25) ..., we *refer to* this gene *as* gip-1 (for gamma tubulin interacting protein) and the corresponding protein *as* CeGrip-1. (JCB2, p.596)

Naming statements are signaled by the main verb such as *refer to*, *name*, *designate*. Given that naming statements follow some important finding by the author, we consider that the ELS class is important, even though we had a small number of examples. In other domains where the methodology is less standardized than in biology (e.g. computer science), there may well be other kinds of information fitting into ELS (e.g. the author's proposal and invention). In that case, further elaboration of the class (e.g. a finer grained ELS class with subclasses NANING and PROPOSAL) would be necessary. We leave this open in the present study.

4.3.9 DIFFERENCE (DFF)

The following are some examples fitting into DFF zones (annotations are ours).

(26) a. [[These effects are distinctly *different* from the effects caused by .....]$_{DFF}$]$_{RSL}$ (PNAS4, p.1925)
b. [[..., we verified that the identified motifs *differ* from the known elements to which PTB binds.]$_{DFF}$]$_{INS}$ (NAR4, pp.1976-77)
c. [[This seems to *be inconsistent with* the view that ...]$_{DFF}$]$_{IMP}$

DFF is signaled by a limited set of vocabulary (mainly, *different* and *contrast* and their variants). Also, as illustrated above, DFF often overlaps with other classes (e.g. INS, IMP, RSL), and therefore involves nested/combined annotation.

4.3.10 CONNECTION (CNN)

The following are some examples fitting into CNN zones (annotations are ours).

(27) a. [[This conservation further *supports* their putative regulatory role in exon skipping.]$_{CNN}$]$_{INS}$ (NAR4, pp.1981-82)
b. [[These data are consistent with the previous report,]$_{CNN}$]$_{RSL}$ [suggesting ...]$_{INS}$
c. [[The results also *confirm* the recent discovery of ... (ref).]$_{CNN}$]$_{INS}$
d. [[This conclusion *was supported* not only by ... but also by ....]$_{CNN}$]$_{INS}$

The CNN class covers statements mentioning consistency (i.e. some kind of positive relation) between data/findings. A generalization of the form is:

(28) a. X pred Y.
pred: is {consistent with/ similar to/ same as}; conforms to; supports
X/Y: previous work, the author's observation, model, hypothesis, insight, etc.
b. X. Similarly, Y. (X/Y: a proposition)

The specific consistency relation varies (e.g. correlation or similarity; support for the author's own or other's data/ idea/ findings). We observed slightly more CNN zones than DFF zones in our sample, and we consider that this is not accidental; this asymmetry indicates that biologists put more focus on correlation between two elements. However, our data regarding these zones is currently too small to support this intuition statistically.

	BKG	PBM	MTH	RSL	INS	IMP	ELS	DFF	CNN	OTL
Lexical (vocabulary)	*	+	*	*	+	+	*	+	+	+
Main verb	*	+	+	+	+	*	+	*	*	+
Tense	*	+	+	+	*	+	+	*	*	+
Sent.-final citation	+	-	-	-	-	-	-	*	*	*
Ref. to Fig./Table	-	-	*	+	-	-	-	*	*	-
Section/location	+	+	+	+	+	+	*	+	+	+
Underlying subject	*	*	+	*	*	*	+	*	*	+

Table 3
Contribution of features in the identification of zones: positive(+); negative(-); irrelevant(∗)

4.4 Summary

Table 3 provides a general picture of the contribution of different feature types to the human identification of zones after the annotation was completed. Because the annotation was done by only one individual at this stage, gathering statistical data related to the degree of contribution of individual features was not possible or meaningful. Rather, our table shows in more general terms which feature types were a positive (+), negative (-), or irrelevant (∗) for the identification process. To take the example of the feature 'sentence-final citation', the row means the following. The presence of a sentence-final citation in the constituent under investigation provides positive evidence for a BKG zone, and negative evidence for the six zone classes including PBM, whereas it is neutral for the last three zone classes including DFF. The table shows, among other facts, that the feature section/location was the most useful positive indication of a zone type, used for identification of all the other zones except for ELS.

Data such as the one provided in this table can be highly useful when moving towards automatic ZA using methods such as machine learning. While the features we have identified in this paper are specifically intended to help human annotation and while defining features ideal for machine learning purposes is a task beyond the scope of this paper, our general summary of the contribution of different feature types for human annotation does provide a valuable starting point for the later purpose of defining such features.

5 Distribution of zones

Another kind of data interesting not only as a general description of the rhetorical structure of biological articles but also for the future purpose of interpreting results of automatic ZA and IE is quantitative data related to the distribution of zones in our 20 annotated articles. In the subsequent sections we first provide a visual illustration of the patterns in which the zones tend to appear (Section 5.1) and then show statistical data illustrating the distribution, nature and location of different zones in articles (Section 5.2).

5.1 Visual illustration

Figures 2 to 5 illustrate the patterns in which the zones appear in the four main sections of the articles. The annotated text is illustrated using different colors for different zone classes, together with the opening/ending tags for each zone. The illustrations exemplify one zone-annotated NAR journal article, but the patterns they highlight are typically observed in the articles we examined. The text is intentionally obscured to focus on the typical zone patterns.

Figure 2 illustrates an annotated I-section. This section has large BKG zones, smaller PBM zones, and a relatively long OTL zone in the end. This is a very typical pattern of the I-section which applies across the articles investigated.

The M-section is, on the other hand, highly homogeneous (Figure 3). With a few exceptions among our annotated articles, the section consists of MTH zones only.

The R-section presents an interesting pattern of zones (Figure 4). We observe a certain motif consisting of a sequence of PBM, MTH, RSL and (optional) INS zones with some variations (e.g. absence of an INS zone; optionally inserted BKG zone etc.). The whole section is a repetition of such motifs. We interpret that the R-section as the problem-solving part of the paper is divided into smaller units which each corresponds to a motif observed.

Figure 5 shows the case of the D-section. The section has as its major elements INS and IMP zones and optional CNN and DFF zones (in this example we see only CNN zones) embedded in INS or IMP zones. This makes sense, given that the section is devoted to a deeper analysis of the results/findings obtained in the present work as well as to prospective remarks.

The quantitative analysis in the following section provides statistical data to complement these observations.

Fig. 2. Zone distribution in the I-section

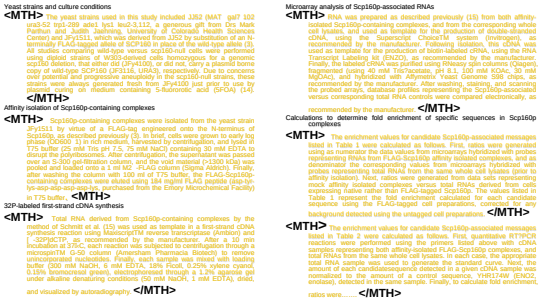

Fig. 3. Zone distribution in the M-section

5.2 Quantitative analysis

Table 4 shows the general distribution of zones in the data of 20 annotated articles consisting of 109,282 words and 3737 sentences. The second row of the table shows the distribution when calculated from the number of words annotated as belonging to each zone (compared to the total number of words in the data). The third and fourth rows show the distribution when calculated from the number of fully annotated sentences (only those sentences annotated with sentence-level annotations) and of all the sentences which contain either constituent-level or sentence-level annotations, respectively.

The figures in this table show that the most frequent zone is RSL; 29% of words

Fig. 4. Zone distribution in the R-section

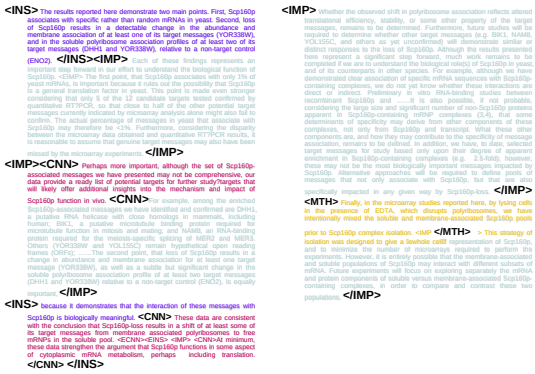

Fig. 5. Zone distribution in the D-section

	BKG	PBM	MTH	RSL	INS	IMP	ELS	DFF	CNN	OTL
Words	20%	5%	26%	**29%**	10%	9%	0%	1%	3%	1%
Full sentences	18%	4%	24%	**26%**	7%	8%	0%	0%	0%	0%
All sentences	20%	8%	28%	**30%**	11%	9%	0%	1%	2%	1%

Table 4
The percentage of words, full sentences and all (including partial) sentences belonging to different zones in the 20 articles. Note that any 0% in this and the following table correspond to some non-zero frequency count.

Sections	BKG	PBM	MTH	RSL	INS	IMP	ELS	DFF	CNN	OTL
I-section	**80%**	9%	2%	3%	4%	0%	0%	0%	0%	1%
M-section	1%	0%	**93%**	0%	1%	2%	0%	0%	0%	1%
R-section	9%	7%	18%	**52%**	10%	5%	0%	2%	3%	0%
D-section	24%	2%	2%	22%	20%	**31%**	0%	1%	5%	2%

Table 5
The percentage of words belonging to different zones in the I-, M-, R- and D-sections of the articles

are annotated as RSL, 26% of full sentences are annotated at the sentence level as RSL, and 30% of all the sentences in the data contain some constituent which is annotated as RSL. MTH is almost as frequent, and the third most frequent zone is BKG. The remaining seven zones are all considerably less frequent. None of them cover more than 10% of words or 8% - 11% of sentences in the data.

Table 5 complements our discussion in the previous section, illustrating the location and distribution of zones in the main sections of the 20 articles: the I-, M-, R- and D-sections. The figures were calculated from the number of words annotated as belonging to each zone in the data. Note that because the zones often overlap, the percentage of words belonging to each zone is independent of the percentages of words belonging to the other zones. According to the table, the most homogeneous sections in terms of zones are indeed the M-section and the I-section. 93% of the M-section belongs to the MTH zone, while 80% of the I-section belongs to the BKG zone. The R-section also has one clearly predominating zone, namely the RSL zone (52%). The most diverse section in term of zones is the D-section, where four different zone classes show up with a similar frequency: BKG, RSL, INS, and IMP (covering 24%, 22%, 20% and 31% of the section, respectively).

The zones which are frequent in more than just one of these four sections are BKG, MTH, INS, and RSL. This shows that in order to focus on the RSL zones (experimental results), for example, we need to look at not only the R-section but also other sections.

The least frequent zones are OTL, ELS, DFF and CNN, none of which contribute more than 5% of any section. However, these zones are not regarded as insignificant; quite the opposite. Specifically, the small percentage of OTL and ELS zones makes sense, given that OTL provides an outline of the paper whereas ELS is committed to 'any other information' within OWN.

Length of zones	BKG	PBM	MTH	RSL	INS	IMP	ELS	DFF	CNN	OTL
Sentence level	31	24	27	29	28	30	16	**37**	30	28
Constituent level	24	14	27	22	26	22	23	24	25	**31**
Blocks (word count)	102	35	156	87	39	71	16	37	52	**392**
Blocks (sent. count)	3.3	1.4	5.8	3.0	1.4	2.4	1.0	1.0	1.8	**14**
Sentences	**31**	19	26	29	26	29	19	26	**31**	27

Table 6
The average length of (i) zones for sentence level and constituent level annotations, (ii) blocks of sentences (unbroken sequences of full sentence annotations for zones), and (iii) sentences containing zone annotations. (All the lengths are indicated by the average number of words, unless otherwise indicated in the table).

Table 6 provides more detailed statistical information about the different zones. The second and third rows show the average length of zones, measured by the number of words, for sentence level and constituent level annotations. For sentence level annotations, the longest average zone is DFF (37 words) and the shortest is ELS (16 words). For constituent level annotations the longest zone is CNN (31 words) and the shortest PBM (14 words). The third and fourth rows of the table show the average length of sentence blocks, i.e. unbroken sequences of full sentence annotations for zones, as measured by the (i) number of words and (ii) the number of sentences belonging to each block. This gives an idea of how zone annotations are scattered and distributed in texts. As expected, the infrequent OTL zone shows the longest average block length (392 words and 14 sentences). This conforms to our observation that OTL zones are usually annotated for a sequence of sentences as a whole. In general, the most frequent zones BKG, MTH and RSL tend to occur in relatively long blocks. Of the three, MTH has the longest blocks (156 words and 5.8 sentences on average). This is explained by the fact that the M-section consists almost exclusively of MTH zones. If we leave the most infrequent zones out of consideration, the most scattered zones are PBM and INS, which typically occur in the blocks of 35-39 words and 1.4 sentences.

Finally, the last row of the table shows the average sentence length in each zone. According to the tables, the zones with the longest sentences are BKG and CNN. The average sentence length in both is 31 words. The shortest sentences can be found in PBM and ELS (19 words). The average sentence length in the articles in general is 29 words.

The data presented in this section shows tendencies that are important to consider for automatic ZA. Issues such as the frequency, location, length, com-

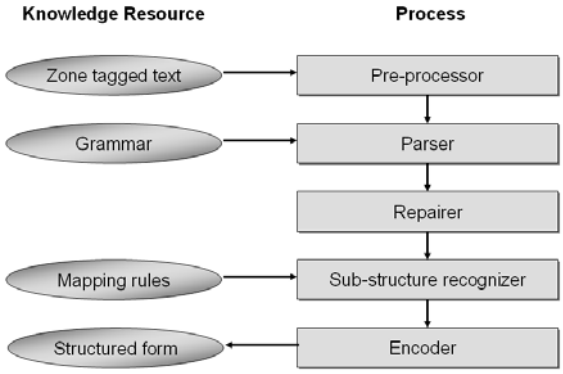

Fig. 6. Preparation of data for machine learning used for ZA

plexity, and distribution of zones have a significant impact on the results of machine learning and should help explain why some zones are more challenging to identify than others.

6 Ongoing and future work

In this paper, we have described our zone annotation scheme for organizing experimental results in biological texts and provided a qualitative and quantitative analysis of the process and the results of ZA based on our hand-annotated sample of 20 articles. This is the first major step towards what is our ultimate goal: to use automatic ZA for the purpose of IE from biological data.

Using our annotated data as training material, we have recently started machine learning experiments for automatic ZA. We have used standard text processing techniques to clean up and convert the annotated data into structured XML format and to extract a variety of basic features for sentences (e.g. n-grams of word tokens and lemmatized words, morphological information about main verbs, and syntactic dependencies). Syntactic features are derived using the Conexor FDG parser [16].

Figure 6 shows a general overview of the processes and knowledge components used for the preparation of training data for this task. The first component, the pre-processor, shown in the figure is responsible for removing any non-zone XML elements from the annotated texts and for preparing XML attributes in a form acceptable to a parser. The parser identifies syntactically interesting relationships in sentences. The repairer corrects wrongly segmented sentences which the parser sometimes outputs. It also tokenizes certain multi-word ex-

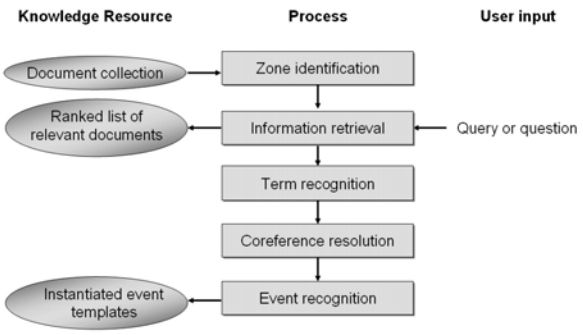

Fig. 7. Using automatic ZA for IR and IE

pressions such as *in_vitro*. Finally, the encoder outputs the text and features in a structured format suitable for machine learning. Certain machine learning experiments are planned with a larger collection of annotated data, using a more comprehensive and refined set of task-specific features, with the view to incorporating the improved process into IE.

An overview of how ZA could be integrated into an information access system is shown in Figure 7. The figure shows how initially ZA would be used in the indexing stage of abstracts and full journal articles to enhance retrieval by scientists. The enhanced information provided by ZA about the rhetorical status of results would allow for more focused querying in a number of scenarios such as locating supporting or conflicting results. This would be for example indicated by the presence of a RSL zone overlapping with a CNN or DFF zone. The returned documents from the retrieval stage could be given directly to the user or, more likely, given to an IE system for more detailed analysis. By having rhetorical zones explicitly annotated, the IE system will be able to focus on identifying instances of relations that meet the needs of the scientist within his/her experimental context: For example, identifying a newly discovered function of a particular entity (cf. in (29)a) or checking to see whether someone has speculated on the cooperation between certain biological entities (cf. in (29)b).

(29) a. ...it is not known if these ligands trigger any signals through $p75^{NTR}$. Thus, our findings demonstrating that $p75^{NTR}$ is a signal transducer not only for neurotrophins but also for ..... (JCB1, p.565)
b. This is the first report demonstrating functional cooperation between a restriction endonuclease and an exonuclease... (NAR3, p.1893)

The IE process itself follows the traditional segmentation of IE into terminology identification/classification, coreference resolution to find identity relations between terms, definite descriptions and pronouns and then finally event extraction. In an advanced IE system (not shown in Figure 6) we envis-

age that sources of information from multiple documents would be integrated to obtain the most timely and reliable results or to place a particular result within the context of a flow of work by the research community. Such a system would make use of citations that link papers, together with a knowledge of how those citations are mentioned in relation to the query.

We are now planning to create more zone annotated data in a systematic way so that it can be used to develop robust machine learning tools. Construction of a larger collection of will be based on the scheme described in this paper and will measure inter- and intra-annotator agreement. This is crucial in order to confirm our intuitions about the stability of the scheme. As a possibility, we are considering to use the ontology management tool developed by us (Open Ontology Forge, http://research.nii.ac.jp/~collier/resources/OOF/index.htm) for the following purposes; 1) to define zone classes as ontology classes (zone annotation is then expected to be a variant of named entity annotation, which we are familiar with), and 2) to link between expressions referring to results (e.g. *these results*; *our results*) and the corresponding RSL zone(s) providing a concrete description of the experimental results), using the coreference tool. Applications include full color representation of annotated texts; a sample is available at: http://research.nii.ac.jp/~collier/projects/ZAISA/index.htm.

7 Conclusions

In this paper, we have focused on the problem that current IE techniques do not provide sufficient or effective means for managing the factual data (particularly data pertaining to experimental results) in the rapidly growing field of biology. We have proposed addressing this problem by means of rhetorical zone analysis. As the first step towards this goal, we have introduced an annotation scheme for biological texts, provided data annotated according to this scheme, and given a comprehensive analysis of the characteristics of this data, useful for both theoretical and practical purposes. We have also described our preliminary research on automatic ZA, and discussed how it could be incorporated into our IE framework. Although our current focus is on biology, we expect that our approach can be applied to a wider area of biomedicine.

Acknowledgements

We gratefully acknowledge the helpful comments from the anonymous reviewers of earlier versions of the paper, and from Patrick Ruth, Udo Hahn and others in the audience of the JNLPBA Workshop held in conjunction with the

COLING 2004 conference. We also thank Simone Teufel (University of Cambridge) and Noriko Kando (NII) for stimulating discussions. Thanks also go to the generous support of Prof. Asao Fujiyama (NII) and the partial financial support from the BioPortal Project performed through Special Coordination Funds for Promoting Science and Technology from the Ministry of Education, Culture, Sports, Science and Technology, the Japanese Government.

We acknowledge the permission by the Oxford University Press (and also by the authors in the case of the NAR1 article) for the citations from the NAR journal articles which are specified in the 'Source online journal articles' section below.

Appendix: Sample zone analysis compared with RST analysis

We contrast our zone analysis with RST analysis using the sample passage below taken from the R-section of a NAR article (NAR1, p.1834). (Tags are inserted for the illustration of RST analysis: the numeral tags in square brackets such as [1] refer to the text span continuing to the next tag, and alphanumeric ones such as [1a] are used at a lower level with the relevant text included.)

> [1][$_{1a}$To address the question of sequence specificity,] [$_{1b}$ RNA samples derived from FLAG-Scp160p-containing complexes versus total RNA from the same cell lysates were used as templates] [$_{1c}$ to generate probes for hybridization to Affymetrix YG-S98 yeast gene chips (see Materials and Methods).] [2] [$_{2a}$As a control,] [$_{2b}$ corresponding pools of RNA derived from cells expressing native, rather than FLAG-tagged Scp160p, also were prepared.] [3] [$_{3a}$The results,] [$_{3b}$ determined by comparing the hybridization results of each Scp160p complex-derived sample against its corresponding total RNA control (see Materials and Methods),] [$_{3c}$ were striking.] [4] First, [$_{4a}$ although many strong hybridization spots were detected in both test and control samples,] [$_{4b}$ the patterns were different,] [$_{4c}$ indicating that the Scp160p complex-derived samples did not simply contain a random subset of total cellular mRNAs.] [5] Furthermore, those sequences most abundant in the mock-isolated samples were completely distinct from those most abundant in the FLAG-Scp160p complex-derived samples (data not shown), [$_{5a}$ demonstrating specificity of the isolation procedure.] [6] In sum, of the 6000 putative yeast gene sequences interrogated on the microarrays in duplicate experiments, only 1% (69 sequences) appeared 2.5-fold enriched in the FLAG-Scp160p complex-derived samples in both experiments (Table 1).

Figure 8 illustrates a RST analysis of the passage. Discourse relations between elements in the passage are represented in a hierarchical manner using various labels such as 'purpose', 'evaluation', and 'interpretation'. The same passage

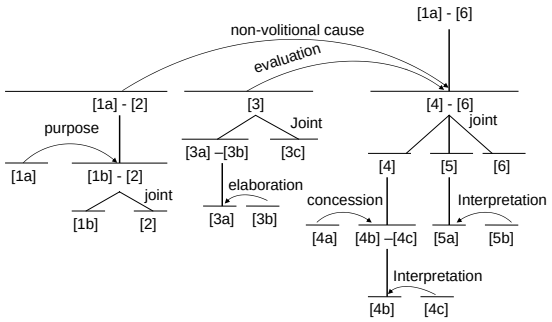

Fig. 8. A sample RST analysis of a biological text

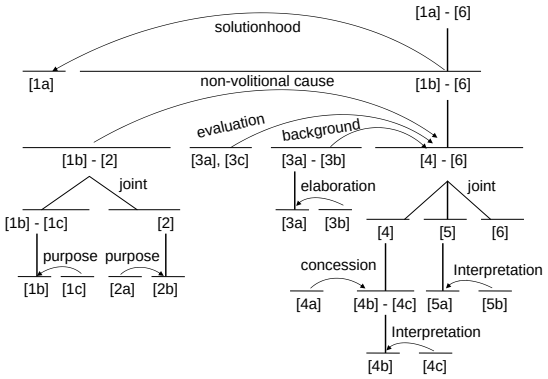

Fig. 9. An alternative RST analysis of the same text (cf. Figure 8)

can be given an alternative RST analysis, which is illustrated in Figure 9.

We mention just one point here. Figure 8 shows an analysis that elements [1a] through [2] describing the purpose of the experiment and the experimental procedure (i.e. *To address the question of ..., RNA samples derived from ...*) serve as a 'non-volitional cause' of the experimental results described by the elements [4] through [6]. In contrast, Figure 9 shows an analysis that element [1a] describing the purpose of the experiment gets a 'solution(hood)' from the rest of the passage. Either interpretation seems to be possible. This ambiguity, besides the rather complex format of the analysis, is likely to cause (inter- and intra-)annotator disagreement.

⟨**PBM**⟩ [1a] ⟨**/PBM**⟩ ⟨**MTH**⟩ [1b] [2a] [2b] ⟨**/MTH**⟩ ⟨**RSL**⟩ [3a] [3b] [4a] [4b] ⟨**/RSL**⟩ ⟨**INS**⟩ [4c] ⟨**/INS**⟩ ⟨**RSL**⟩ [5] [6] ⟨**/RSL**⟩

Fig. 10. Zone analysis of the same text (cf. Figures 8 and 9)

Figure 10 illustrates the ZA. The text is divided into zones employed in our scheme, as indicated here by the inserted opening/ending tags in boldface which specify the zone boundaries as well as the corresponding zone class. For example, the span of text [1b] through [2b] between the opening tag ⟨**MTH**⟩ and the ending tag ⟨**/MTH**⟩ (i.e. 'RNA samples derived from ..... also were prepared') is annotated as a MTH zone. Compared with the two versions of RST analysis illustrated in Figures 8 and 9, the ZA demonstrates a much simpler structure and yet captures the kinds of information to our concern. This is expected to contribute to a higher (inter- and intra-) annotator agreement and a better identification of the information to our IE needs.

Source online journal articles

NAR Neuclid Acid Research Online (URL http://nar.oxfordjournals.org), Oxford University Press.

NAR1 A-M Li, A. Watson and J. L. Fridovich-Keil, Scp160p associates with specific mRNAs in yeast, *NAR* 31(7), 2003, pp.1830–1837.

NAR3 N.K. Raghavendra and D. N. Rao, Functional cooperation between exonucleases and endonucleases-basis for the evolution of restriction enzymes, *NAR* 31(7), 2003, pp.1888–1896.

NAR4 E. Miriami, H. Margalit and R. Sperling, Conserved sequence elements associated with exon skipping, *NAR* 31(7), 2003, pp.1974–1983.

NAR6 N. Fujikawa, H. Kurumizaka, O. Nureki, T. Terada, M. Shirouzu, T. Katayama and S. Yokoyama, Structural basis of replication origin recognition by the DnaA protein, *NAR* 31(8), 2003, pp.2077–2086.

PNAS Proceedings of the National Academy of Sciences Online (URL http://www.pnas.org), The National Academy of Sciences of the United States of America.

PNAS3 I. G. Minko and Y. Zou and R.S. Lloyd, Biochemistry incision of DNA-protein crosslinks by UvrABC nuclease suggests a potential repair pathway involving nucleotide excision repair, *PNAS* 99(4), 2002, pp.1905–1909.

PNAS4 M. V. Kumar and T. Shimokawa, T. R. Nagy, and M. D. Lane, Differential effects of a centrally acting fatty acid synthase inhibitor in lean and obese mice, *PNAS* 99(4), 2002, pp.1921–1925.

PNAS5 M. Uno and K. Ito and Y.Nakamura, Polypeptide release at sense and noncognate stop codons by localized charge-exchange alterations in translational release factors, *PNAS* 99(4), 2002, pp.1819–1824.

JCB Journal of Cell Biology Online (URL http://www.jcb.org), The Rockefeller University Press.

JCB1 T. Yamashita, H. Higuchi and M. Tohyama, The p75 receptor transduces the signal from myelin-associated glycoprotein to Rho, *JCB*, 157(4), 2002, pp.565–570.

JCB2 E. Hannak, K. Oegema, M. Kirkham, P. Gōnczy, B. Habermann and A. A. Hyman, The kinetically dominant assembly pathway for centrosomal asters in Caenorhabditis elegans is γ-tubulin dependent, *JCB*, 157(4), 2002, pp.591–602.

References

[1] M. Craven, J. Kumlien, Constructing biological knowledge bases by extracting information from text sources, in: Proceedings of the 7th International Conference on Intelligent Systemps for Molecular Biology (ISMB-99), Heidelburg, Germany, 1999, pp. 77–86.

[2] K. Humphreys, G. Demetriou, R. Gaizauskas, Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures, in: Proceedings of the Pacific Symposium on Biocomputing (PSB2000), 2000, pp. 505–516.

[3] L. Tanabe, W. Wilbur, Tagging gene and protein names in biomedical text, Bioinformatics 18 (2002) 1124–1132.

[4] Y. Mizuta, N. Collier, An annotation scheme for a rhetorical analysis of biology articles, in: Proceedings of the Fourth International Conference on Language and Evaluation (LREC2004), Lisbon, Portugal, 2004, pp. 1737–1740.

[5] S. Teufel, M. Moens, Summarizing scientific articles: Experiments with relevance and rhetorical status, Computational Linguistics 28 (4) (2002) 409–445.

[6] Y. Mizuta, N. Collier, Zone identification in biology articles as a basis for information extraction, in: Proceedings of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA) at the COLING2004 International Conference, Geneva, Switzerland, 2004, pp. 29–35.

[7] J. Swales, Genre analysis, Cambridge University Press, 1990.

[8] E. Liddy, The discourse-level structure of empirical abstracts: An explanatory study, Information Processing and Management 27 (1) (1991) 55–81.

[9] N. Kando, Text structure analysis as a tool to make retrieved document usable, in: Proceedings of the 4th International Workshop on Information Retrieval with Asian Languages, Taipei, Taiwan, 1999, pp. 126–135.

[10] W. Mann, S. Thompson, Rhetorical structure theory: Toward a functional theory of text organization, Text 8 (3) (1987) 243–281.

[11] R. Prasad, E. Miltsakaki, A. Joshi, B. Webber, Annotation and data mining of the penn discourse treebank, in: Proceedings of the ACL Workshop on Discourse Annotation (downloadable at http://www.cis.upenn.edu/~pdtb/papers/acl-discourse-annotation.pdf, last visited on April 7, 2005), Barcelona, Spain, 2004.

[12] A. Lehman, Text structuration leading to an automatic summary system, Information Processing and Management 35 (2) (1999) 181–191.

[13] S. Teufel, H. van Halteren, Agreement in human factoid annotation for summarization evaluation, in: Proceedings of the Fourth International Conference on Language and Evaluation (LREC2004), 2004.

[14] L. Carlson, D. Marcu, M. E. Okurowski, Building a discourse-tagged corpus in the framework of rhetorical structure theory, in: Proceedings of the 2nd SIGDIAL Workshop on Discourse and Dialogue (downloadable at http://acl.ldc.upenn.edu/W/W01/W01-1605.pdf, last visited on April 7, 2005), Aalborg, Denmark, 2001.

[15] Y. Mizuta, T. Mullen, N. Collier, Annotation of biomedical texts for zone analysis, Tech. Rep. NII-2004-007E, National Institute of Informatics, ISSN 1346-5597 (2004).

[16] P. Tapanainen, T. Järvinen, A non-projective dependency parser, in: Proceedings of the Fifth Conference on Applied Natural Language Processing, Washington Marriot Hotel, Washington D.C., Association of Computational Linguistics, 1997, pp. 64–71.